Lecture 8 introduces a classification scheme for protein folds and the protein shape universe.

Topic 1.
Why are there relatively few unique protein folds, while the sequence landscape is so vast? Speculate on the success rate of de novo protein designs (i.e., the problem of designing a protein fold that is unique among known crystallographic structures). Comment on what lessons were learned from the literature on an example of a de novo protein. Are ideas of recurrent domains and domain fusion useful for protein engineering structures with new function? Explain.

A big reason as to why there are a few unique protein folds is that essentially, the purpose of protein folding is to place hydrophobic residues in the protein's core, away from water or a solvent. Additionally, it has been hypothesized that proteins fold in the lowest energy state that their amino acid composition can access. With such limiting factors in addition to the slow speed at which evolution works, the number of unique protein folds are many magnitudes smaller than the amount of unique protein sequences. This makes sense as protein folds serve a functional purpose, where if it works correctly, it is conserved until a fold comes along that works better. I believe that a huge part in why protein folds are so heavily conserved can be related to the adage "if it ain't broke, don't fix it".

In de novo protein design, neither the sequence nor the structure of the protein are known. To start with, different structural conformations are tested and "made" using small peptide fragments, which are then run through sequence optimization to figure out if this newly designed structure is the lowest-energy state of the sequence. If approaching de novo protein design without any restrictions, there are far too many combinations and conformations to test, but when limiting it with specific rules, it becomes a bit more manageable. Some important constraints to adhere to when performing de novo protein design are those that are sequence-independent. These sequence-independent constraints are important in developing structural geometry because they cut down on the structural permutations available for design. A few of these constraints are that the polar atoms in the structural backbone must be in contact with hydrogen bonds in the chain or with the solvent via exposed loops. Another constraint is to take in the amount of flexibility that a polypeptide chain has, thus restricting the lengths of loops that can connect alpha helices and beta sheets in various orientations[1].

Domain fusion is when a protein in a given species appears to be the fusion of two different proteins in another species. Through the fusion of these two separate proteins, the entropy of dissociation gets reduced and it is hypothesized that the two proteins shared a function or a physical interaction with each other[2]. If applied to de novo protein design, I believe that domain fusion could play a huge role in developing new proteins that are more efficient than current proteins by combining two functionally related proteins into a sort of "super protein" that achieves functionality for both.

Sources:

1. **Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. *Nature*. 2016;537(7620):320-327. doi:10.1038/nature19946.**
2. **Chia JM, Kolatkar PR. Implications for domain fusion protein-protein interactions based on structural information. *BMC Bioinformatics*. 2004;5:161. Published 2004 Oct 26. doi:10.1186/1471-2105-5-161**

Topic 2.
Using the SCOP library and its hierarchical scheme, find a superfamily of all-beta proteins that contain five or more families. Report the fold type, superfamily and families.
Superfamily: P-53-like Transcription Factors
Fold type: Common fold of diphtheria toxin/transcription factors/cytochrome f
Families:

1. **p53 DNA-binding domain-like**
2. **Rel/Dorsal transcription factors, DNA-binding domain**
3. **T-box**
4. **STAT DNA-binding domain**
5. **RUNT domain**
6. **DNA-binding domain from NDT80**
7. **DNA-binding protein LAG-1 (CSL)**

Due date is 04 Nov